

CLEARINGHOUSE FOR FEDERAL SCIENTIFIC AND TECHNICAL INFORMATION CFSTI  
DOCUMENT MANAGEMENT BRANCH 410.11

LIMITATIONS IN REPRODUCTION QUALITY

ACCESSION # 410.11-106

- 1. WE REGRET THAT LEGIBILITY OF THIS DOCUMENT IS IN PART UNSATISFACTORY. REPRODUCTION HAS BEEN MADE FROM BEST AVAILABLE COPY.
- 2. A PORTION OF THE ORIGINAL DOCUMENT CONTAINS FINE DETAIL WHICH MAY MAKE READING OF PHOTOCOPY DIFFICULT.
- 3. THE ORIGINAL DOCUMENT CONTAINS COLOR, BUT DISTRIBUTION COPIES ARE AVAILABLE IN BLACK-AND-WHITE REPRODUCTION ONLY.
- 4. THE INITIAL DISTRIBUTION COPIES CONTAIN COLOR WHICH WILL BE SHOWN IN BLACK-AND-WHITE WHEN IT IS NECESSARY TO REPRINT.
- 5. LIMITED SUPPLY ON HAND: WHEN EXHAUSTED, DOCUMENT WILL BE AVAILABLE IN MICROFICHE ONLY.
- 6. LIMITED SUPPLY ON HAND: WHEN EXHAUSTED DOCUMENT WILL NOT BE AVAILABLE.
- 7. DOCUMENT IS AVAILABLE IN MICROFICHE ONLY.
- 8. DOCUMENT AVAILABLE ON LOAN FROM CFSTI ( TT DOCUMENTS ONLY).
- 9.

NBS 9/64

PROCESSOR: 11

604206 604206

ON SOME VARIATIONAL PROBLEMS OCCURRING  
IN THE THEORY OF DYNAMIC PROGRAMMING

Richard Bellman  
Irving Glicksberg  
Oliver Gross

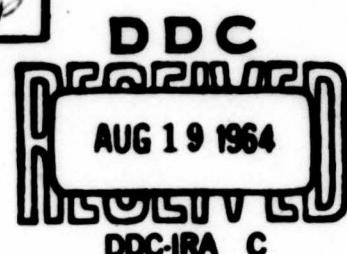
P-380 ✓

1 May 1953

Approved for OTS release

51 p

COPY	OF	1
HARD COPY		\$3.00
MICROFICHE		\$0.50 p



The RAND Corporation  
SANTA MONICA • CALIFORNIA

ON SOME VARIATIONAL PROBLEMS OCCURRING IN  
THE THEORY OF DYNAMIC PROGRAMMING

Richard Bellman  
Irving Glicksberg  
Oliver Gross

§1. Introduction.

The purpose of this paper is to present some results of an investigation of a class of interesting and important variational problems involving the control of a physical system over a time interval. One large category of problems of this nature arises in connection with the maintenance of a dynamic system in or near a specified state at minimum cost. The cost is usually compounded of two parts, the first part measured in terms of the deviation of the system from the desired state, and the second part measured by the cost of the resources used for control. The theory of mechanical, electrical, and economic systems contain many questions of this type.

Another large category of problems, of economic and industrial origin, are those in which it is required to maximize the output of a system given a limited quantity of resources. Only one representative of this category will be discussed in this paper.

The mathematical difficulties encountered in treating problems of the types above depend to a large degree upon the mathematical model used to represent the system, the functionals employed to measure the cost of deviation and the cost of resources, and the constraints imposed upon the permissible types of control.

As far as the mathematical models are concerned, we shall consider here physical systems which are ruled by a system of linear differential equations of the form

$$(1) \quad \frac{dx_1}{dt} = \sum_{j=1}^N a_{1j} x_j + f_1(t), \quad i=1,2,\dots,N$$

$$x_1(0) = c_1, \quad 0 \leq t \leq T$$

which in the more convenient vector-matrix notation takes the form

$$(2) \quad \frac{dx}{dt} = Ax + f(t), \quad x(0) = c, \quad 0 \leq t \leq T$$

where

$$(3) \quad x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{pmatrix}, \quad f(t) = \begin{pmatrix} f_1(t) \\ f_2(t) \\ \vdots \\ f_N(t) \end{pmatrix}, \quad c = \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_N \end{pmatrix}.$$

Here the vector  $x$  represents the state of the system at any time  $t$ , and the forcing term  $f(t)$  represents the influence of the resources used for control.

If, in place of continuous control, we apply intermittent control, the equation above is replaced by a difference equation,

$$(4) \quad x(t+1) = Ax(t) + f(t), \quad t=0,1,2,\dots,T,$$

$$x(0) = c.$$

If we consider systems in which time lags occur, in place of (2) or (4) we obtain equations of the form

$$(5) \quad \frac{dx(t)}{dt} = Ax(t) + Bx(t-1) + f(t), \quad 1 \leq t \leq T$$

$$x = \phi(t), \quad 0 \leq t \leq 1.$$

A discussion of control problems involving differential-difference equations will be postponed to a subsequent paper because of the variety of new features which arise.

As far as the functionals representing the cost are concerned, we shall take them to be linear or quadratic, with constraints of linear form involving boundedness and positivity.

We discuss first, in §2, the case where the cost of deviation is measured by  $\int_0^T [x-y, x-y] dt$ , where  $[u, v]$  represents the inner product of two vectors, where  $x$  is the actual state,  $y$  the desired state, and the cost of control is measured by a  $\int_0^T [f, f] dt$ . The problem is then that of choosing an  $f$  so as to minimize the total cost.

$$(6) \quad J(f) = \int_0^T [x-y, x-y] dt + a \int_0^T [f, f] dt$$

where  $x$  depends upon  $f$  by way of (2). Using the known representation theorems expressing the solutions of inhomogeneous linear differential equations in terms of the solutions of the homogeneous equations and the forcing terms, the problem may be reduced to

minimizing the functional

$$(7) \quad \int_0^T \left[ \int_0^t K(t, t_1) f(t_1) dt_1 - y, \int_0^t K(t, t_1) f(t_1) dt_1 - y \right] dt_1 + \int_0^T [f, f] dt$$

over all  $f$ .

A completely analogous problem is obtained for difference equations in which integrals are replaced by sums. For the case of differential-difference equations a functional corresponding to (7) is obtained.

To treat these various problems simultaneously we abstract the general problem and consider the problem of minimizing

$$(8) \quad \|Af + g\|^2 + a\|f\|^2$$

where  $A$  is a bounded linear operator on functions belonging to  $L^2(0, T)$ , and  $\|f\|^2 = \int_0^T [f, f] dt = (f, f)$ .

Related to the problem of minimizing the total cost is that of minimizing the cost of deviation subject to the restriction that the cost of control be bounded by a fixed quantity, and the dual problem of minimizing the cost of control given that the cost of deviation is to remain below a certain bound.

After having obtained the existence and uniqueness of the minimizing function in the general case, together with the integral

equation it satisfies, we turn to the application of the general result to the differential equation model.

In the next three sections we consider some particular problems in which the cost functions are of a varied type and in which there are constraints on the forcing term  $f$ . In §3 we treat the problem of minimizing  $\int_0^T (1-u)^2 dt$  over all  $f$  subject to

$$(9) \quad \begin{aligned} (a) \quad 0 &\leq f \leq m, \\ (b) \quad \int_0^T f dt &\leq a_1, \end{aligned}$$

where  $u$  is related to  $f$  by means of the equation

$$(10) \quad \frac{du}{dt} = -u + f, \quad u(0) = 1.$$

In the succeeding section, §4, we treat the problem of minimizing  $\int_0^T (du/dt)^2 dt$  subject to the same constraints.

In §5 we treat the problem of minimizing  $\max_{0 \leq t \leq T} |1-u|$  subject to  $\int_0^T f^2 dt \leq a_1$ , by considering the solution of the dual problem of minimizing  $\int_0^T f^2 dt$  subject to  $-d \leq 1-u \leq d$  for  $0 \leq t \leq T$ .

Turning from this class of problems we consider a problem arising in the mathematical theory of economics concerning maximization of profit. Setting

$$(11) \quad \frac{dx_1}{dt} = \sum_{j=1}^N a_{1j} y_j, \quad x_1(0) = c_1$$

where  $a_{ij} \geq 0$ , the problem is that of choosing  $y_1$  between 0 and  $x_1$  so as to maximize the functional

$$(12) \quad \int_0^T \left[ \sum_{j=1}^N (x_j - y_j) \right] dt.$$

In the course of the solution it is necessary to answer the following question, of some independent interest. Given

$$(13) \quad \frac{dx_1}{dt} = \sum_{j=1}^N a_{1j} x_j + f_1(t), \quad x_1(0) = c_1$$

what are the necessary and sufficient conditions upon the matrix  $A = (a_{ij})$  in order that  $x_1 \geq 0$  for  $t \geq 0$  whenever  $c_1 \geq 0$  and  $f_1(t) \geq 0$ ?

The answer turns out to be quite simple, namely  $a_{1j} \geq 0$ ,  $i \neq j$ .

There are a multitude of interesting questions which we have not mentioned at all. A quite important one is that where there are two forcing terms,

$$(14) \quad \frac{dx}{dt} = Ax + f + g$$

the first,  $f$ , representing factors at our control, and the second,  $g$ , representing exogenous factors beyond our control. Although we shall not discuss any problems of this type in this paper, let us merely point out that there are at least two approaches to the corresponding minimization problems. We may regard  $g$  as a random function with known expected value and autocorrelation function,

and minimize  $E(g(f,g))$ . Or we may introduce the concepts of game theory and consider the problem of determining the  $f$  and  $g$  which yield

$$(15) \quad \min_f \max_g J(f,g), \quad \max_g \min_f J(f,g),$$

and then consider the corresponding game over function space.

Finally we mention that there are several alternative approaches to the problems we discuss. If we allow the problem to remain in its native form (6), subject to differential side condition, we have the classical problem of Bolza. It seems, however, simpler to consider the space of functions  $f \in L^2(0,T)$  than the space of functions possessing derivatives.

In this connection we can also use the Lagrange multiplier approach, particularly in the problems of the later sections. We have preferred to use a direct attack based upon the Neyman-Pearson Lemma. Subsequently we will present a new approach to a more extensive class of problems based upon dynamic programming techniques.

## §2. Quadratic Functional.

In this section let us consider the problem of minimizing

$$(1) \quad J_a(f) = (g + Af, g + Af) + a(f,f), \quad a > 0,$$

where  $f$  and  $g$  belong to  $L^2(0,T)$  and  $A$  is a bounded linear operator on  $L^2(0,T)$ . The function  $J_a(f)$  represents, from the above models, the total cost of control, where the first term represents the cost of deviation and the second term the cost of control.

Theorem 1. There is a unique  $f \in L^2(0,T)$  which furnishes the minimum to  $J_a(f)$ .

Proof: Since  $\sqrt{J_a(f)}$  may be interpreted as the norm of the element  $[g + Af, f]$  of  $L^2 \times L^2$ , the existence and uniqueness of the minimizing  $f$ , which we shall call  $f_a$ , is a simple consequence of the fact that a strongly closed convex set in Hilbert space has a unique element of minimal norm. It is clearly sufficient to consider those  $f$  for which  $a(f, f) \leq \inf_f J_a(f)$ . These form a weakly compact convex set. Since  $A$  is weakly continuous, the image of this set in  $L_2 \times L_2$  is weakly closed, hence strongly closed, and convex.

Theorem 2. Let

$$(2) \quad R_{-a} = (-a - A^*A)^{-1}$$

be the resolvent operator of the positive operator  $A^*A$  (where  $A^*$  is, as usual, the adjoint operator to  $A$ ), then

$$(3) \quad f_a = R_{-a} A^*g.$$

Proof: Let  $\lambda$  be an arbitrary real constant,  $f$  an arbitrary function in  $L^2(0,T)$  and consider

$$(4) \quad \begin{aligned} J(f_a + \lambda f) &= (g + Af_a + \lambda Af, g + Af_a + \lambda Af) + a(f_a + \lambda f, f_a + \lambda f) \\ &= J_a(f_a) + 2\lambda [ (g + Af_a, Af) + a(f_a, f) ] \\ &\quad + \lambda^2 [ (Af, Af) + a(f, f) ]. \end{aligned}$$

The condition that  $J(f_a + \lambda f)$  be a minimum at  $\lambda = 0$  yields

$$(5) \quad 0 = (g + Af_a, Af) + a(f_a, f) \\ = (A^*g + A^*Af_a + af_a, f),$$

for all  $f$ , which implies

$$(6) \quad A^*g + A^*Af_a + af_a = 0,$$

which in turn yields (3), since  $-a < 0$  is in the resolvent set of the positive operator  $A^*A$ .

Now, as is well known,  $R_\lambda$  is an analytic function in the resolvent set,  $\frac{d}{d\lambda} R_\lambda = -R_\lambda^2$ ,  $\frac{d}{d\lambda} R_\lambda^2 = -2R_\lambda^3$ , and

$R = -\lambda^{-1}(1 + \lambda^{-1}A^*A)^{-1}$  tends uniformly to zero as  $\lambda \rightarrow \infty$ . Thus  $(R_{-a}^2 A^*g, A^*g) = \|f_a\|^2 \rightarrow 0$  as  $a \rightarrow \infty$ , and  $\|f_a\|^2$  is non-increasing as a function of  $a > 0$ , for

$$(7) \quad \frac{d}{da} \|f_a\|^2 = \frac{d}{da} (R_{-a}^2 A^*g, A^*g) = 2(R_{-a}^3 A^*g, A^*g) \\ = 2(R_{-a} f_a, f_a) \leq 0$$

since  $R_{-a}$ , as the inverse of a strictly negative operator, is strictly negative; indeed since  $f_a = 0$  only if  $A^*g = 0$ ,  $\|f_a\|^2$  is strictly decreasing if  $A^*g \neq 0$ . Similarly, since

$$\begin{aligned}
 (12) \quad (g + Af_a, g + Af_a) &= (g + Af_a, g) + (A^*g + A^*Af_a, f_a) \\
 &= (g, g) + (f_a, A^*g) - a(f_a, f_a). \\
 &= (g, g) + (R_{-a} A^*g, A^*g) - a(f_a, f_a),
 \end{aligned}$$

we have

$$\begin{aligned}
 (13) \quad \frac{d}{da} \|g + Af_a\|^2 &= \frac{d}{da} \left[ (R_{-a} A^*g, A^*g) - a(R_{-a}^2 A^*g, A^*g) \right] \\
 &= (R_{-a}^2 A^*g, A^*g) - (R_{-a}^2 A^*g, A^*g) \\
 &\quad - 2a(R_{-a}^3 A^*g, A^*g) \\
 &= -2a(R_{-a} f_a, f_a) \geq 0.
 \end{aligned}$$

Thus  $\|g + Af_a\|^2$ , as a function of  $a > 0$ , is non-decreasing, and strictly increasing if  $A^*g \neq 0$ .

These properties enable us to easily dispense with the related problems of minimizing the cost of deviation with a limited amount of control, that is, obtaining

$$(10) \quad \phi(c) = \min_{\|f\|^2 \leq c} \|g + Af\|^2,$$

---

\* The monotone character of  $\|f_a\|^2$  and  $\|g + Af_a\|^2$  can also be seen for  $A^*g \neq 0$  from the fact that  $f_a \neq f_b$ ,  $0 < a < b$ . For if  $u_a = \|g + Af_a\|^2$ ,  $v_a = \|f_a\|^2$ , then  $u_a + av_a < u_b + bv_b$  and  $u_b + bv_b < u_a + bv_a$  so that  $u_b + bv_b < u_a + av_a + (b-a)v_a < u_b + av_b + (b-a)v_a$  and  $(b-a)v_b < (b-a)v_a$  or  $v_b < v_a$ ; adding  $-av_a < -av_b$  to  $u_a + av_a < u_b + bv_b$  we obtain  $u_a < u_b$ .

and the dual problem of obtaining

$$(11) \quad \Psi(c) = \min_{\|g+Af\|^2 \leq c} \|f\|^2,$$

which is the minimum cost of control required to keep the cost of deviation below a certain level.

Let us consider first the trivial case where  $A^*g = 0 = f_a$  (for  $a > 0$ ). Since  $J_a(f_a) = \|g\|^2 \leq J_a(f) = \|g+Af\|^2 + a\|f\|^2$  for any  $f$  and all  $a > 0$ ,  $\|g\|^2 \leq \|g+Af\|^2$  for all  $f$  and clearly  $\phi(c) = \|g\|^2$  for all  $c \geq 0$ . Moreover we have  $\Psi(c) = 0$  for  $c \geq \|g\|^2$ , with  $\Psi(c)$  undefined for  $c < \|g\|^2$ . In the somewhat less trivial case of  $A^*g \neq 0$ , since  $\|f_a\|^2$  is continuous and strictly decreasing, for  $c$  in the range  $0 < c < \sup \|f_a\|^2 = \lim_{a \rightarrow 0} \|f_a\|^2$ , we clearly have a unique  $a > 0$  for which  $\|f_a\|^2 = c$ , and thus  $f_a$  alone provides  $\phi(c)$ , since otherwise we should have an  $f \neq f_a$  for which  $\|g+Af\|^2 \leq \|g+Af_a\|^2$ ,  $\|f\|^2 \leq \|f_a\|^2$  so that  $J_a(f) \leq J_a(f_a)$ , which is impossible. Thus, to complete our discussion of (10), we need only consider the case in which  $\sup \|f_a\|^2$  is finite, and  $c > \sup \|f_a\|^2$ .

If  $\sup \|f_a\|^2 = \lim_{a \rightarrow 0} \|f_a\|^2$  is finite, then there is an element  $f_0$  to which  $f_a$  converges strongly as  $a \rightarrow 0$ , which minimizes  $\|g+Af\|^2$  for all  $f$  in  $L_2$ , hence provides  $\phi(c)$  for all  $c \geq \sup \|f_a\|^2$ . This follows from the fact that the set  $\{f_a\}$  has a weak cluster point  $f_0$  for which  $\|f_0\|^2 \leq \lim_{a \rightarrow 0} \|f_a\|^2$ , and the minimal property of  $\|g+Af_a\|$ . For since  $g+Af_0$  is a cluster point of  $g+Af_a$ , we have  $\|g+Af_0\| \leq \lim_{a \rightarrow 0} \|g+Af_a\| \leq \|g+Af_a\|$ . Hence, if  $\|g+Af_0\| < \|g+Af_a\|$ , we have

$J_a(f) < J_a(f_a)$  for sufficiently small  $a$ , contradicting the minimal property. Thus,  $\|g + Af_0\| \leq \|g + Af\|$  for all  $f$ . Furthermore,  $\|g + Af\| = \|g + Af_0\|$  if and only if  $Af = Af_0$ , from the strict convexity of the unit sphere in  $L_2$ . Now if  $Af = Af_0$  and therefore  $\|g + Af\| \leq \|g + Af_a\|$  for all  $a$ , we must have  $\|f\| > \|f_a\|$  for all  $a$  so that  $\|f\|^2 \geq \lim_{a \rightarrow 0} \|f_a\|^2 \geq \|f_0\|^2 \geq \lim_{a \rightarrow 0} \|f_a\|^2$ . Thus,  $f_0$  is the unique element of minimal norm in the closed variety  $\{f: Af = Af_0\}$ .

Since this is true for any weak cluster point  $f_0$ , there is only one and since  $\|f_a\| \rightarrow \|f_0\|$ ,  $\|f_0 - f_a\| \rightarrow 0$  as asserted. Also, since  $A^*g + A^*Af_a = -af_a$  tends strongly to zero, we obtain  $A^*g + A^*Af_0 = 0$ .

Similar considerations apply to (11) when  $A^*g \neq 0$ . For  $c$  in the range of  $\|g + Af_a\|^2$ ,  $0 < a < \infty$ , we have a unique  $a$  for which  $\|g + Af_a\|^2 = c$  and  $f_a$  alone provides  $\Psi(c)$ . If  $c$  is not in this range, we have two cases. If  $c < \|g + Af_a\|^2$  for all  $a > 0$  and  $\|f_a\|^2$  is unbounded, the problem is vacuous, since  $\|g + Af\|^2 \leq c < \|g + Af_a\|^2$  implies  $J_a(f) < J_a(f_a)$  for  $\|f_a\|^2 > \|f\|^2$ . But if  $\sup \|f_a\|^2 < \infty$ , then since the element  $f_0$  provides the absolute minimum of  $\|g + Af\|^2$ ,  $\Psi(c)$  is defined only if  $c = \|g + Af_0\|^2$ , and then  $\Psi(c) = \|f_0\|^2$ . On the other hand, if  $c > \|g + Af_a\|^2$  for all  $a > 0$ , then since  $\|f_a\|^2 \rightarrow 0$  as  $a \rightarrow \infty$ ,  $c \geq \|g\|^2$ , and  $\Psi(c) = 0$  is provided by  $f = 0$ .

Thus

Theorem 3. If  $A^*g \neq 0$ , either  $c = \|f_a\|^2$  for some  $a > 0$ , in which case  $f_a$  alone provides  $\Psi(c)$  or  $c > \|f_a\|^2$  and  $f_0 = \lim_{a \rightarrow \infty} f_a$  provides the minimum. For  $c$  in the range of  $\|f_a\|^2$ ,

$$(12) \quad \frac{d\phi(c)}{dc} = -a$$

where  $c = \|f_a\|^2$  relates  $c$  and  $a$ .

Similarly, either  $c = \|g+Af_a\|^2$  for some  $a > 0$ , in which case  $f_a$  alone provides  $\psi(c)$ , or  $\|f_a\|^2$  is bounded and  $c = \|g+Af_0\|^2$ , in which case  $f_0$  alone provides our minimum  $\psi(c)$ , or  $c > \|g+Af_a\|^2$  for all  $a > 0$ , and  $f = 0$  provides the minimum. Also, in the range of  $\|g+Af_a\|^2$ ,

$$(13) \quad \frac{d\psi(c)}{dc} = -\frac{1}{a}, \quad c = \|g+Af_a\|^2.$$

It only remains to verify (12) and (13). Since  $\frac{d\phi(c)}{dc} = \frac{d}{da} \|g+Af_a\|^2 \frac{da}{dc} = -2a(R_{-a} f_a, f_a) \frac{da}{dc}$  and  $c = (f_a, f_a) = (R_{-a}^2 A^* g, A^* g)$ ,

$$(14) \quad 1 = 2(R_{-a}^2 A^* g, A^* g) \frac{da}{dc} = 2(R_{-a} f_a, f_a) \frac{da}{dc},$$

and (12) holds. Similarly, for  $c = \|g+Af_a\|^2$ ,  $1 = -2a(R_{-a} f_a, f_a) \frac{da}{dc}$  and

$$(15) \quad \frac{d\psi(c)}{dc} = \frac{a}{dc} + \|f_a\|^2 = 2(R_{-a} f_a, f_a) \frac{da}{dc},$$

and (13) holds.

Let us finally observe that if for some element  $h \in L^2(\omega, \mathbb{C})$  we have  $g = -Ah$ , then the value of  $J_a(f_a)$  is given by

$$\begin{aligned}
 (16) \quad J_a(f_a) &= (g + Af_a, g + Af_a) + a(f_a, f_a) \\
 &= (g + Af_a, g) + (g + Af_a, Af_a) + a(f_a, f_a) \\
 &= (g + Af_a, g),
 \end{aligned}$$

by virtue of (5). Using  $g = -Ah$  and (5) again, we obtain

$$(17) \quad J_a(f_a) = a(f_a, h)$$

### §3. Application to Differential Equations.

Let us consider the application of the preceding results to the case where the system under control is ruled by a set of differential equations of the form

$$(1) \quad \frac{dx_1}{dt} = \sum_{j=1}^N b_{1j}x_j + f_1(t),$$

$$x_1(0) = c_1, \quad i=1, 2, \dots, N,$$

or by an  $n$ -th order equation

$$\begin{aligned}
 (2) \quad \frac{d^N u}{dt^N} + a_1 \frac{d^{N-1} u}{dt^{N-1}} + \dots + a_N u &= f(t), \\
 u^{(k)}(0) &= c_k, \quad k=0, 1, 2, \dots, N-1.
 \end{aligned}$$

In the latter case we consider the problem of minimizing

$$(3) \quad J(f) = \int_0^T \left[ \sum_{k=0}^{N-1} b_k \left( \frac{d^k u}{dt^k} - c_k \right)^2 \right] dt + a \int_0^T f^2 dt,$$

while in the former we wish to minimize

$$(4) \quad J(f) = \int_0^T \left[ \sum_{k=1}^N b_k (x_k - c_k)^2 \right] dt + a \int_0^T \left[ \sum_{k=1}^N f_k^2 \right] dt.$$

Since every  $N$ -th order linear equation may be converted into an  $N$ -th order linear system by means of the substitution

$$(5) \quad x_1 = u$$

$$x_2 = \frac{du}{dt}$$

$$\vdots \quad d^{N-1}u$$

$$x_N = \frac{d^{N-1}u}{dt^{N-1}},$$

we shall confine our attention to systems. These are most effectively discussed using vector matrix technique. Set

$$(6) \quad x = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{pmatrix}, \quad f(t) = \begin{pmatrix} f_1(t) \\ f_2(t) \\ \vdots \\ f_N(t) \end{pmatrix}, \quad c = \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_N \end{pmatrix},$$

$$B = (a_{ij}), \quad i, j = 1, 2, \dots, N$$

Equation (1) may now be written

$$(7) \quad \frac{dx}{dt} = Bx + f(t), \quad x(0) = c.$$

Let us assume for simplicity that the coefficients  $b_k$  in (4) are all unity, and use the usual inner product notation

$$(8) \quad [x, y] = \sum_{i=1}^N x_i y_i.$$

The expression to be minimized takes the form

$$(9) \quad J(f) = \int_0^T [x-c, x-c] dt + a \int_0^T [f, f] dt.$$

Furthermore, we define the norm as

$$(10) \quad \|f\| = (f, f)^{1/2} = \left( \int_0^T [f, f] dt \right)^{1/2}.$$

To convert this problem into the type discussed in § 2, we require the following well-known result in the theory of linear differential equations.

Lemma 1. The solution of (7) may be written in the form

$$(11) \quad x = y + \int_0^t y(t-t_1) f(t_1) dt_1,$$

where y is the solution of the homogeneous equation

$$(12) \quad \frac{dy}{dt} = By, \quad y(0) = c,$$

and  $Y$  is the matrix solution of

$$(13) \quad \frac{dy}{dt} = BY, \quad Y(0) = I,$$

which is to say  $Y = e^{Bt}$ ,  $y = e^{Bt}c$ .

If we set

$$(14) \quad g = y - c$$

$$Af = \int_0^t Y(t-t_1)f(t_1)dt_1,$$

then  $x = g + Af$ , and the variational problem is now a special case of that considered previously.

Furthermore, since  $y$  satisfies (12), we have

$$(15) \quad \frac{dg}{dt} = Bg + Bc, \quad g(0) = 0,$$

whence

$$(16) \quad g(t) = \int_0^t Y(t-t_1)Bc dt_1 = ABC(t).$$

The adjoint operator to  $A$ ,  $A^*$ , is defined by

$$(17) \quad A^*f = \int_t^T Y(t-t_1)f(t_1)dt_1.$$

We obtain this by considering

$$\begin{aligned}
 (18) \quad (\mathbf{A}f, g) &= \int_0^T [\mathbf{A}f, g] dt, \\
 &= \int_0^T \left[ \int_0^t Y(t-t_1) f(t_1) dt_1, g(t) \right] dt. \\
 &= \int_0^T \left[ f(t_1), \int_{t_1}^T Y(t-t_1)' g(t) dt \right] dt_1,
 \end{aligned}$$

as we see by interchanging the orders of integration, where  $Y(t-t_1)'$  is the transpose of  $Y(t-t_1)$ , the matrix  $e^{B'(t-t_1)}$ , where  $B'$  is the transpose of  $B$ .

Referring to the previous section, the minimizing  $f$  is given by

$$(19) \quad f = (\mathbf{I} - \mathbf{a} - \mathbf{A}^* \mathbf{A})^{-1} \mathbf{A}^* g,$$

which means that  $f$  satisfies the integral equation,

$$(20) \quad \mathbf{a}f + \mathbf{A}^* \mathbf{A}f = \mathbf{A}^* g.$$

From (17) and (20) we obtain the condition

$$(21) \quad f(T) = 0.$$

Since the inverse of  $\mathbf{A}^*$  is  $\frac{d}{dt} - B'$ , we obtain from (20)

$$(22) \quad \mathbf{a} \left( \frac{df}{dt} - B' f \right) + \mathbf{A}f = -g,$$

which means that

$$(23) \quad \frac{df}{dt} - B'f = 0 \text{ at } t = 0.$$

Using the operator  $\frac{d}{dt} - B$  we have

$$(24) \quad a\left(\frac{d}{dt} - B'\right)\left(\frac{d}{dt} - B\right)f + f = -\left(\frac{d}{dt} - B\right)g = -Bc.$$

This is a system of linear differential equations subject to the two-point boundary conditions of (21) and (23). The solution exists and is unique by virtue of Theorem 1.

#### §4. Application to Difference Equations.

Similar results hold for the variational problem associated with the system of difference equations

$$(1) \quad x(t+1) = Ax(t) + f(t), \quad t=0,1,2,\dots,T$$

$$x(0) = c,$$

with the norm defined by

$$(2) \quad \|f\| = \left( \sum_{t=0}^T [f(t), f(t)] \right)^{\frac{1}{2}}.$$

The analogue of Lemma 1 is

Lemma 2. The solution of (1) may be written

$$(3) \quad x(t) = y(t) + \sum_{t_1=0}^t y(t-t_1-1)f(t_1),$$

where

$$(4) \quad \begin{aligned} y(t+1) &= By(t), & y(0) &= c, \\ Y(t+1) &= BY(t), & Y(0) &= I, \end{aligned}$$

which is to say  $Y = B^t$ ,  $y = B^t c$ .

The remaining details are now completely analogous.

### §5. Differential-difference Equations.

If we consider problems of continuous control with a time lag we meet functional equations of the form

$$(1) \quad \begin{aligned} \frac{du(t)}{dt} &= au(t) + bu(t-1) + f(t), & t > 1 \\ u(t) &= g(t), & 0 \leq t \leq 1. \end{aligned}$$

Although results similar to the above hold, we shall postpone discussion of these until a later paper devoted solely to equations of this type, since some additional difficulties arise.

### §6. A Result Concerning Positivity.

Let us agree to call a vector  $x$  non-negative if all of its components are non-negative and write  $x \geq 0$ , and similarly call a matrix  $A$  non-negative if  $a_{ij} \geq 0$ , writing again  $A \geq 0$ .

Using this notation we shall prove

Theorem 4. The necessary and sufficient condition that the solution of

$$(1) \quad \frac{dx}{dt} = Ax + f(t), \quad x(0) = c,$$

be non-negative for  $t \geq 0$  whenever  $f(t)$  and  $c$  are non-negative  
is that

$$(2) \quad a_{ij} \geq 0, \quad i \neq j.$$

Proof: The solution of (1) has the form

$$(3) \quad x = e^{At} c + \int_0^t e^{A(t-t_1)} f(t_1) dt_1,$$

from which it follows that if  $x \geq 0$  for all  $c \geq 0$  and  $f(t) \geq 0$ , we must have  $e^{At} \geq 0$ , and clearly this is sufficient.

The problem then reduces to finding the necessary and sufficient condition that  $e^{At} \geq 0$  for  $t \geq 0$ . Since  $e^{At} = I + At + \dots$ , it is clear that  $a_{ij} \geq 0$ ,  $i \neq j$  is necessary in order that  $e^{At} \geq 0$  for small positive  $t$ . The following simple proof that this condition is sufficient is due to S. Karlin. We have  $e^{At} = (e^{At/n})^n$  for any integer  $n$ . Choosing  $n$  large enough, we will have  $e^{At/n} \geq 0$  for  $0 \leq t \leq t_0$ , by virtue of  $a_{ij} \geq 0$ . Since the product of non-negative matrices is non-negative, we obtain the desired result.

In the case of variable  $A(t)$  sufficiency at least may be established readily by means of the change of variable

$$(4) \quad \begin{aligned} & \int_0^t a_{11}(s) ds \quad y_1, \quad i=1,2,\dots,N, \\ & x_1 = e \end{aligned}$$

converts (1) into the form

$$(5) \quad \frac{dy_1}{dt} = \sum_{j \neq 1} a_{1j} y_j + r_1(t) e^{- \int_0^t a_{11}(t) dt},$$

$$y_1(0) = c_1.$$

The sufficiency of the condition  $a_{1j}(t) \geq 0, \quad i \neq j$ , is now clear.

### §7. A Problem in Mathematical Economics.

Let us consider the following idealized problem in mathematical economics. We have a system with  $N$  outputs measured by the variables  $x_i(t)$ ,  $i=1,2,\dots,N$ . Each output  $x_i$  is divided into two parts  $y_i$  and  $z_i$  where  $z_i$  is taken out as profit and  $y_i$  is reinvested to increase future output. Assuming that the change in output is determined by the equations

$$(1) \quad \frac{dx_i(t)}{dt} = \sum_{j=1}^N a_{ij} y_j(t), \quad i=1,2,\dots,N,$$

$$x_i(0) = c_i,$$

with  $a_{ij} \geq 0$ ,

what reinvestment policy does one follow in order to maximize the total profit,

$$\int_0^T \sum_{i=1}^N (x_i - y_i) dt ?$$

Since it is not difficult to establish the existence of a solution, we shall omit this point and turn immediately to obtaining the solution.

In order to illustrate clearly the techniques involved, we shall treat in succession, the one-dimensional, two-dimensional, and N-dimensional problem.

The One-dimensional Problem.

We have

$$(2) \quad \frac{dx_1}{dt} = a_{11}y_1, \quad a_{11} > 0, \quad x_1(0) = c_1$$

and we wish to maximize

$$(3) \quad J_1 = \int_0^T (c_1 + a_{11} \int_0^t y_1 dt_1 - y_1) dt,$$

where  $y_1$  is subject to the conditions

$$(5) \quad 0 \leq y_1 \leq c_1 + a_{11} \int_0^t y_1 dt.$$

An interchange in the order of integration in (4) yields

$$(6) \quad J_1 = c_1 T + \int_0^T (a_{11}(T-t)-1)y_1 dt_1.$$

Let  $T_1$  be the value of  $t$  for which

$$(7) \quad a_{11}(T-t)-1 = 0,$$

assuming for the moment that  $T > 1/a_{11}$ . The function  $y_1$  which maximizes (6) subject to (5) is then given by

$$(8) \quad y_1 = c_1 + a_{11} \int_0^t y_1 dt, \quad 0 \leq t \leq T_1,$$

$$= 0 \quad , \quad T_1 \leq t \leq T$$

Note that  $T_1$  depends upon  $T$ . If  $T \leq 1/a_{11}$ ,  $y_1=0$  is the maximizing function. There is no difficulty in obtaining the explicit form of  $y_1$ .

The Two-dimensional Problem.

Consider the problem of maximizing

$$(9) \quad J_2 = \int_0^T (z_1 + z_2) dt,$$

where

$$(10) \quad \frac{dx_1}{dt} = a_{11}y_1 + a_{12}y_2, \quad i=1,2,$$

$$x_1(0) = c_1,$$

$$a_{ij} \geq 0, \text{ and, finally, } z_i = x_i - y_i, \quad 0 \leq y_i \leq x_i.$$

Solving for the  $x_i$  in terms of the  $y_i$  in (10) we obtain

$$(11) \quad x_1 = c_1 + a_{11} \int_0^t y_1 dt + a_{12} \int_0^t y_2 dt, \quad i=1,2.$$

The expression for  $J_2$  then takes the form

$$\begin{aligned}
 (12) \quad J_2 &= \int_0^T \left[ c_1 + a_{11} \int_0^t y_1 dt + a_{21} \int_0^t y_1 dt - y_1 \right] dt \\
 &= \int_0^T \left[ c_2 + a_{12} \int_0^t y_2 dt + a_{22} \int_0^t y_2 dt - y_2 \right] dt \\
 &= (c_1 + c_2)T + \int_0^T \left( (a_{11} + a_{21})(T-t) - 1 \right) y_1 dt \\
 &\quad + \int_0^T \left( (a_{12} + a_{22})(T-t) - 1 \right) y_2 dt.
 \end{aligned}$$

Let  $T_1, T_2$  be given by

$$(13) \quad (a_{11} + a_{21})(T - T_1) - 1 = 0,$$

$$(a_{12} + a_{22})(T - T_2) - 1 = 0,$$

and take  $T$  large enough so that  $T_1$  and  $T_2$  are positive. Assume without loss of generality further that  $T_2 > T_1$ .

A partial solution to our maximization problem is then given by

$$(14) \quad y_1 = y_2 = 0, \quad T_2 \leq t \leq T$$

$$z_2 = 0, \quad 0 \leq t \leq T_2,$$

$$z_1 = 0, \quad 0 \leq t \leq T_1.$$

The only unknown remaining is the value of  $z_1$  in  $T_1 \leq t \leq T_2$ . Returning to the expression for  $J_2$  in (9) and using the partial results of (14) we obtain

$$(15) \quad J_2 = \int_{T_1}^{T_2} z_1 dt_1 + (T-T_2) (x_1(T_2) + x_2(T_2)) .$$

Employing (9) we obtain

$$(16) \quad \begin{aligned} x_1(T_2) &= c_1 + a_{11} \int_0^{T_2} y_1 dt + a_{12} \int_0^{T_2} y_2 dt \\ &= c_3 + a_{11} \int_{T_1}^{T_2} y_2 dt + a_{12} \int_{T_1}^{T_2} x_2 dt, \end{aligned}$$

where  $c_3$  is a constant independent of the value of  $y_2$  in  $[T_1, T_2]$ , and similarly

$$(17) \quad x_2(T_2) = c_4 + a_{21} \int_{T_1}^{T_2} y_1 dt + a_{22} \int_{T_1}^{T_2} x_2 dt.$$

Using  $z_1 = x_1 - y_1 = c_1 + a_{11} \int_0^t y_1 dt + a_{12} \int_0^t y_2 dt - y_1$ , we obtain finally

$$(18) \quad \begin{aligned} J_2 &= c_5 + \int_{T_1}^{T_2} (c_6(T-t)-1) y_1 dt + a_{12} \left( \int_{T_1}^{T_2} \int_0^t y_2 dt \right) dt \\ &\quad + (a_{12} + a_{21}) \int_{T_1}^{T_2} x_2 dt. \end{aligned}$$

To proceed further, we require an expression for  $\int_0^t y_2 dt$  for  $0 \leq t \leq T_2$ . In this interval we have

$$(19) \quad \frac{d}{dt} \left( \int_0^t y_2 dt \right) = y_2 = c_2 + a_{21} \int_0^t y_1 dt + a_{22} \int_0^t y_2 dt,$$

and thus solving, we obtain

$$(20) \quad \int_0^t y_2 dt = e^{a_{22}t} \int_0^t e^{-a_{22}s} c_2 + a_{21} \int_0^s y_1 dt_1 ds \\ = \phi_1(t) + a_{21} e^{a_{22}t} \int_{T_1}^t \left( e^{-a_{22}s} \int_{T_1}^s y_1(t_1) dt_1 \right) ds,$$

for  $T_1 \leq t \leq T_2$ , where  $\phi_1$  is independent of the value of  $y_1$  in  $[T_1, T_2]$ . Hence

$$(21) \quad \int_{T_1}^{T_2} x_2 dt = \int_{T_1}^{T_2} y_2 dt \\ = c_1 + a_{21} e^{a_{22}T_2} \int_{T_1}^{T_2} \left( e^{-a_{22}s} \int_{T_1}^s y_1(t_1) dt_1 \right) ds.$$

Interchanging orders of integration, this is

$$(22) \quad c_1 + a_{21} e^{a_{22}T_2} \int_{T_1}^{T_2} \left( y_1(t_1) \int_{t_1}^{T_2} e^{-a_{22}s} ds \right) dt_1.$$

The important point to observe is that the coefficient of  $y_1$ , namely

$a_{21} e^{a_{22}T_2} \int_{t_1}^{T_2} e^{-a_{22}s} ds$ , is a decreasing function of  $t_1$ .

It remains to simplify the expression  $\int_{T_1}^{T_2} \left( \int_0^t y_2 dt \right) dt$ .

We have

$$(23) \int_{T_1}^{T_2} \left( \int_0^t y_2 dt \right) dt = c_8 + a_{21} \int_{T_1}^{T_2} \left[ e^{a_{22}t} \int_{T_1}^t \left( e^{-a_{22}s} \int_{T_1}^s y_1(t_1) dt_1 \right) ds \right] dt$$

$$= c_8 + a_{21} \int_{T_1}^{T_2} \left[ e^{a_{22}t} \int_{T_1}^t y_1(t_1) \left( \int_{t_1}^t e^{-a_{22}s} ds \right) dt_1 \right] dt.$$

The integral has the form

$$(24) \int_{T_1}^{T_2} e^{a_{22}t} \left( \int_{T_1}^t y_1(t_1) \psi(t, t_1) dt_1 \right) dt$$

$$= \int_{T_1}^{T_2} y_1(t_1) \left( \int_{t_1}^{T_2} e^{a_{22}t} \psi(t, t_1) dt \right) dt_1.$$

We have

$$(25) \frac{d}{dt} \left( \int_{t_1}^{T_2} e^{a_{22}t} \psi(t, t_1) dt \right) = -e^{a_{22}t_1} \psi(t_1, t_1)$$

$$+ \int_{t_1}^{T_2} e^{a_{22}t} \frac{\partial \psi}{\partial t_1} dt$$

$$= 0 + \int_{t_1}^{T_2} e^{a_{22}t} \left( -e^{-a_{22}t_1} \right) dt_1 < 0.$$

Hence the coefficient of  $y_1$  in (24) is monotone decreasing. Referring to (18) and observing that  $c_6(T-t)-1$  is monotone decreasing, since  $c_6 > 0$ , we see that the total coefficient of  $y_1$  will be decreasing in  $(T_1, T_2)$  when  $J_2$  is written in the form

$$(26) J_2 = c_8 + \int_{T_1}^{T_2} k(t_1) y_1(t_1) dt_1.$$

To maximize then, we choose  $y_1$  as large as possible in  $[T_1, T_3]$  where  $k \geq 0$ , and equal to zero in  $[T_3, T_2]$ , where  $k(T_3) = 0$ .

Since  $c_6(T-t)-1$  is negative for  $t > T_1$  and the other coefficients are zero at  $T_2$ , it follows that  $T_3$  is actually between  $T_1$  and  $T_2$ .

We have thus demonstrated that the maximum of  $J_2$  subject to (10) et seq. is given by

$$(27) \quad \begin{aligned} y_1 &= y_2 = 0, & T_2 \leq t \leq T \\ z_2 &= 0, & 0 \leq t \leq T_2 \\ z_1 &= 0, & 0 \leq t \leq T_3, \end{aligned}$$

where  $T_3$  is a definite number between  $T_2$  and  $T_2$ , and

$$(28) \quad \begin{aligned} (a_{11} + a_{21})(T-T_1)-1 &= 0, \\ (a_{12} + a_{22})(T-T_2)-1 &= 0, \end{aligned}$$

for  $T > 1/(a_{11} + a_{21}) > 1/(a_{12} + a_{22})$ .

The other cases admit of similar solutions. The conditions  $a_{11} \geq 0$  may be relaxed to  $a_{11} + a_{21} \geq 0$ ,  $a_{12} + a_{22} \geq 0$ .

#### The N-dimensional Problem.

If we examine the details of the previous case we see that everything hinges on the fact that  $e^{a_{22}t}$  is non-negative. In order to see what the required analogue is, let us consider the N-dimensional case using vector-matrix notation.

We have, as before

$$(29) \quad J = \sum_{j=1}^N \int_0^T \left[ \left( \sum_{i=1}^N a_{ij} \right) (T-t)-1 \right] y_j dt,$$

where  $0 \leq y_j \leq x_j$  and the  $x_j$  satisfy (1). Taking  $T$  large enough, let  $0 < T_1 < T_2 < \dots < T_N$  be given by

$$(30) \quad \left( \sum_{i=1}^N a_{ij} \right) (T - T_j) - 1 = 0.$$

As above, it follows immediately that  $y_N$  is given by

$$(31) \quad \begin{aligned} y_N &= x_N, & 0 \leq t \leq T_N, \\ &= 0, & T_N < t \leq T. \end{aligned}$$

We may then eliminate  $y_N$  and solve for  $y_{N-1}$  in  $(T_{N-1}, T_N)$  the only interval in which it is unknown.

At the very next step, when eliminating  $y_{N-1}$  and  $y_N$  and expressing them in terms of the other  $y_k$  we are confronted by the problem of solving a system of equations of the form

$$(32) \quad \frac{du_i}{dt} = \sum_{j=R+1}^N a_{ij} u_j + \sum_{j=1}^R a_{ij} \int_0^t y_j dt + c_i, \quad i=R+1, \dots, N$$

for the  $u_i$ ,  $i=R+1, \dots, N$ , in terms of the  $y_j$ ,  $j=1, 2, \dots, R$ , and of determining the monotonicity properties of the coefficients of the  $y_j$ .

In order to solve this problem we employ vector-matrix notation. Let

$$(33) \quad v = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{pmatrix}, \quad b = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{pmatrix}, \quad A = (a_{ij}), \quad i, j = 1, 2, \dots, n,$$

and consider the system

$$(34) \quad \frac{dv}{dt} = Av + b \int_0^t y_k dt, \quad v(0) = 0.$$

where all the components in  $b$  are non-negative. The expression for  $v$  is

$$(35) \quad v = e^{At} \int_0^t e^{-As} b \left( \int_0^s y_k dt \right) ds.$$

Interchange of order of integration yields

$$(36) \quad v = e^{At} \int_0^t \left( \int_{t_1}^t e^{-As} ds \right) b y_k(t_1) dt_1.$$

The matrix function  $e^{At} \int_{t_1}^T e^{-As} ds$  enters as one coefficient.

The derivative with respect to  $t_1$  is  $-e^{A(T-t_1)}$ .

Similarly, when we consider  $\int_0^T v dt$  we obtain

$$(37) \quad \int_0^T e^{At} \left( \int_0^t \psi(t, t_1) b y_k(t_1) dt_1 \right) dt$$

$$= \int_0^T \left( \int_{t_1}^T e^{At} \psi(t, t_1) dt \right) b y_k(t_1) dt_1.$$

Since

$$(38) \quad \frac{d}{dt_1} \int_{t_1}^T e^{At} \Psi(t, t_1) dt = -e^{At_1} \Psi(t_1, t_1) + \int_{t_1}^T e^{At} \frac{\partial \Psi}{\partial t_1} dt_1$$

$$= 0 - \int_{t_1}^T e^{A(t-t_1)} dt_1,$$

we see that everything depends upon the signs of the elements of  $e^{At}$ . We have, however, in § 6, demonstrated that all the elements will be positive in  $e^{At}$  if  $a_{ij} \geq 0$  for  $i \neq j$ .

Using the above results and the previous techniques, we may establish that the solution to the maximization problem has the same general form for all dimensions, namely,

$$(39) \quad y_k = x_1, \quad 0 \leq t \leq s_k, \\ = 0, \quad s_k < t \leq T, \quad k=1,2,\dots,n.$$

The computation of the numbers  $s_k$  is laborious but straightforward.

Let us observe, finally, that the simplicity of the above result is due to the fact that we assumed all the coefficients were non-negative. Actually, all that is required is that  $a_{ij} \geq 0$ ,  $i \neq j$  and that

$$(40) \quad \sum_{k=1}^n a_{kj} \geq 0, \quad j=1,2,\dots,n.$$

In the general case where the  $a_{ij}$  are both positive and negative, the problem will be much more difficult.

### 68. Quadratic and Linear functionals.

If either the cost of control or the cost of deviation is taken to be a linear functional, and linear constraints of physical origin are introduced, the complexity of the problem of minimizing the total cost is greatly increased. Essentially this is due to the fact that unrestricted variations are in general no longer permissible.

The problem now requires a combination of classical variational techniques and Neyman-Pearson-type techniques blended in an adroit manner.

Our first result is

Theorem 5. Let  $x$  be the absolutely continuous solution on  $[0, T]$  of

$$\frac{dx}{dt} = -x + f, \text{ a.e., } x(0) = 1. \text{ Then the minimum of } \int_0^T (1-x)^2 dt$$

subject to  $\int_0^T f dt \leq a < T, 0 \leq f \leq M (M > 1)$  is furnished by

$$(1) \quad f(t) = \begin{cases} 0 & t < \log(1/1-\lambda) \\ 1-\lambda & \log(1/1-\lambda) \leq t \leq \log(1/1-\lambda) + a/1-\lambda \\ 0 & \log(1/1-\lambda) + a/1-\lambda < t, \end{cases}$$

where  $\lambda$  is determined by a transcendental equation given below.

The minimum of  $\int_0^T (dx/dt)^2 dt$  under the same conditions is furnished by

$$(2) \quad f(t) = \begin{cases} \sqrt{(t-b)^2 + \rho(t-b)} & t \leq b \\ 0 & b \leq t \leq T \end{cases}$$

where  $\mu$ ,  $\rho$ ,  $b$  are determined by transcendental equations, given below.

Proof: Let  $S$  denote the subset of  $L_2(0, T)$  of all  $f$  for which  $0 \leq f \leq M$ ,  $\int_0^T f dt \leq a$ , which is weakly compact. Since

$$(3) \quad x(t) = e^{-t} + e^{-t} \int_0^t e^s f(s) ds,$$

the mappings  $f \rightarrow 1-x$ ,  $f \rightarrow dx/dt = -x + f$  are weakly continuous so that the images are weakly (hence strongly) closed convex sets and have unique elements of minimal norm. Since each mapping is easily seen to be (1-1), in each case there is a unique minimizing  $f$ .

Let  $f_0$  minimize  $J(f) = \int (1-x)^2 dt$ , and for  $f$  in  $S$  let  $f_\lambda = (1-\lambda) f_0 + \lambda f$ ,  $0 \leq \lambda \leq 1$ ,

$$(4) \quad \phi(\lambda) = J(f_\lambda) = \int_0^T (1-e^{-t} - e^{-t} \int_0^t e^s [(1-\lambda)f_0(s) + \lambda f(s)] ds)^2 dt.$$

Since  $\phi(0)$  must be the minimum of  $\phi$  on  $[0, 1]$ , we have

$$(5) \quad 0 \leq \phi'(0) = 2 \int_0^T (1-x_0)(-e^{-t} \int_0^t e^s (f(s) - f_0(s)) ds) dt,$$

where  $x_0(t) = e^{-t} + e^{-t} \int_0^t e^s f_0(s) ds$ . Since clearly  $\phi'' \geq 0$ , this condition implies  $J(f_0) = \phi(0) \leq \phi(1) = J(f)$ . Thus  $f_0$  is the unique element of  $S$  for which

$$(6) \quad \int_0^T (1-x_0)e^{-t} \int_0^t e^s f_0(s) ds dt \geq \int_0^T (1-x_0)e^{-t} \int_0^t e^s f(s) ds dt$$

for all  $f$  in  $S$ . Interchange of the order of integration yields

$$(7) \quad \int_0^T f_0(s) \left[ e^s \int_s^T e^{-t} (1-x_0) dt \right] ds \geq \int_0^T f(s) \left[ e^s \int_s^T e^{-t} (1-x_0) dt \right] ds$$

so that  $f_0$  maximizes  $(f, K_0) = \int_0^T f K_0 ds$ , over  $S$ , where

$$(8) \quad K_0(s) = e^s \int_s^T e^{-t} (1-x_0) dt$$

(which of course depends on  $f_0$ ), and the determination of  $f_0$  appears as a problem of the Neyman-Pearson type.

Before we pursue  $f_0$  further, note that for  $\alpha \geq 1$ ,  $f \leq \alpha$  (a.e) implies  $x \leq \alpha$ , with strict inequality for  $\alpha > 1$  (in particular  $x(t) < M$ ); for  $e^t f_0(t) = d/dt(e^t x_0(t)) \leq \alpha e^t$  and thus  $e^t x_0(t) - 1 \leq \alpha e^t - \alpha \leq \alpha e^t - 1$ . Also  $K_0(t) = \alpha$  on a set implies, as one sees by differentiation, that  $x_0(t) = 1 - \alpha = f_0(t)$  (a.e) on this set.\*

With these simple facts in mind we can now deduce several facts about  $K_0$  which will determine  $f_0$ . First  $E = \{t: K_0(t) > 0\}$  is non-void; otherwise, since clearly  $f_0(t) = 0$  (a.e) where  $K_0(t) < 0$  and  $f_0(t) = 1 - 0 = 1$  where  $K_0(t) = 0$ , we should have  $x_0(t) \leq 1$  and not identically 1 (since, if  $x_0(t) = 1$ , then  $f_0(t) = 1$  a.e. and  $\int_0^T f_0 dt > a$ ) so that  $K_0(t) = e^t \int_t^T e^{-s} (1-x_0(s)) ds > 0$  for some  $t$  despite the assumption that  $K_0 \leq 0$ .

Secondly, the measure of the non-void set  $E$  (which is open since  $K_0$  is continuous) exceeds  $a/M$ . For if this is not the case those  $f$  in  $S$  which maximize  $(f, K_0)$  have  $f(t) = M$  for  $t \notin E$ ; in particular, since  $f_0(t) = M$  on  $E$ ,  $K_0$  is twice differentiable in  $E$  and  $K_0'(t) = K_0(t) - (1-x_0(t))$ ,  $K_0''(t) = K_0'(t) + x_0'(t) = K_0'(t) + M - x_0(t)$ .

---

\* For  $\alpha e^{-t} = \int_t^T e^{-s} (1-x_0(s)) ds$  and thus  $x_0(t) = 1 - \alpha = e^{-t} + e^{-t} \int_0^t e^s f_0(s) ds$ , so  $(1-\alpha)e^t = e^t f_0(t)$ , a.e.

Consequently  $K_0$  has no maximum on  $E \cap (0, T)$  (at such a maximum  $t$   $K_0(t) = 0$  and  $0 \geq K_0''(t) = M - x_0(t) > 0$ ). Thus  $K_0$  is monotonic on components of  $E$ , and, as is easily seen, if  $t \in E$  then either  $[0, t]$  is contained in  $E$  and  $K_0$  is non-increasing there, or  $[t, T]$  is contained in  $E$  and  $K_0$  is non-decreasing there. The latter cannot be the case since  $K_0(T) = 0$ ; neither can the former, since then  $0 \geq K_0'(0) = K_0(0) - (1 - x_0(0)) = K_0(0) > 0$ , and we come to the contradictory conclusion that  $E$  must be void, so that the measure of  $\{t: K_0(t) > 0\}$ ,  $|\{t: K_0(t) > 0\}| > a/M$ .

Since this is the case, there is a non-void set of  $\mu > 0$  for which  $|\{t: K_0(t) \geq \mu\}| \geq a/M$ ; let  $\lambda$  be the sup of these  $\mu$ . Then  $|\{t: K_0(t) \geq \lambda\}| = |\bigcup_{\mu < \lambda} \{t: K_0(t) \geq \mu\}| \geq a/M$  and  $|\{t: K_0(t) > \lambda\}| \leq a/M$  since  $\{t: K_0(t) > \lambda\}$  is the union of an increasing sequence of sets  $\{t: K_0(t) \geq \mu_n > \lambda\}$  each of which has measure  $< a/M$ . In view of this last fact every  $f$  in  $M$  maximizing  $(f, K_0)$  has value  $M$  in  $\{t: K_0(t) > \lambda\}$  and by exactly the argument used in the previous paragraph we find this set is void.

Thus, setting  $E = \{t: K_0(t) = \lambda = \sup K_0\}$ ,  $|E| > a/M$  and any  $f$  in  $M$  maximizing  $(f, K_0)$  vanishes outside  $E$ . Since  $x_0$  is strictly decreasing where  $f_0$  vanishes and assumes the value  $1 - \lambda$  on  $E$ , the closed set  $E$  must be an interval  $[t_1, t_2]$ , and clearly  $t_1 = \log(1 - \lambda)^{-1}$  since  $x_0(t) = e^{-t}$ ,  $t \leq t_1$ . Now it is obvious that any  $f$  which assumes the value  $M$  on a subset of  $E$  of measure  $a/M$  and zero elsewhere maximizes  $(f, K_0)$ , so that the maximum is  $M \lambda a/M = \lambda a$ ;  $(f_0, K_0) = \lambda a = (1 - \lambda) \lambda |E|$ ,  $|E| = a/(1 - \lambda)$  and we see that  $f_0$  has the form indicated. We obtain an equation for  $\lambda$  from  $K(t_2) = \lambda$ ,  $x(t) = (1 - \lambda)e^{t_2 - t}$  (for  $t \geq t_2$ ) which yields as the equation for  $\lambda$ ,

$$(9) \quad \frac{3}{2}(1-\lambda) = 1 - (1-\lambda)^{-1} e^{\frac{a}{1-\lambda} - T} + \frac{1}{2}(1-\lambda)^{-3} e^{\frac{3a}{1-\lambda} - T}.$$

The minimization of  $J(f) = \int_0^T (dx/dt)^2 dt$  over  $S$  can be

achieved in the same manner, as we shall see. Let  $f_0$  minimize this functional  $J$  and set  $f_\lambda = (1-\lambda)f_0 + \lambda f$ ,  $\phi(\lambda) = J(f_\lambda)$ ,  $0 \leq \lambda \leq 1$  or

$$(10) \quad \phi(\lambda) = \int_0^T ((1-\lambda)f_0(t) + \lambda f(t) - e^{-t} - e^{-t}) \int_0^t e^s [(1-\lambda)f_0(s) + \lambda f(s)] ds)^2 dt$$

for  $f \in S$ . Once again we have

$$(11) \quad 0 \leq \phi'(0) = 2 \int_0^T (f_0 - x_0)(f(t) - f_0(t) - e^{-t}) \int_0^t e^s [f(s) - f_0(s)] ds dt$$

and  $f_0$  as the unique element of  $S$  for which

$$(12) \quad \int_0^T (f_0 - x_0)(e^{-t} \int_0^t e^s f_0(s) ds - f_0(t)) dt \geq \int_0^T (f_0 - x_0)(e^{-t} \int_0^t e^s f(s) ds - f(t)) dt$$

for all  $f$  in  $S$ . Interchange of the order of integration yields

$$(13) \quad \int_0^T f_0(s) \left[ e^s \int_s^T e^{-t} (f_0 - x_0) dt - (f_0(s) - x_0(s)) \right] ds \geq$$

$$\int_0^T f(s) \left[ e^s \int_s^T e^{-t} (f_0 - x_0) dt - (f_0(s) - x_0(s)) \right] ds$$

so that  $f_0$  maximizes the inner product  $(f, K_0)$  where we set

$$(14) \quad K_o(s) = x_o(s) - f_o(s) + e^s \int_s^T e^{-t} (f_o - x_o) dt$$

for all  $s$ .\*  $K_o$  need not be continuous in this case, of course, but  $K_o + f_o$  is, and this may be used to provide the analogues of the previous arguments.

Suppose first that  $K_o(t) \leq 0$  a.e. Since  $f_o$  maximizes  $(f, K_o)$ ,  $f_o(t) = 0$  on all but a subset  $E$  of  $\{t: K_o(t) < 0\}$  of measure zero. If we decrease  $f_o$  to zero on  $E$  (so that by (14) we increase  $K_o$ ) then for the altered and clearly equivalent  $f_o$  and  $K_o$  we have  $f_o(t) = 0$  whenever  $K_o(t) < 0$ . Now for the altered  $K_o$  we have  $\{t|K_o(t) < 0\}$  open, for otherwise we have  $t_n \rightarrow t$ ,  $K_o(t) < 0$  and  $K_o(t_n) = 0$  so that  $K_o(t) + f_o(t) = K_o(t) < 0 \leq K_o(t_n) + f_o(t_n)$ , which would contradict the continuity of  $K_o + f_o$  guaranteed by (14).

Since  $f_o(t) = 0$  on this open set  $K_o$  is continuous and differentiable on it, and

$$\begin{aligned} K'_o(t) &= x'_o(t) + e^t \int_t^T e^{-s} (f_o - x_o) ds - (f_o(t) - x_o(t)) \\ &= x'_o(t) + K_o(t) = f_o(t) - x_o(t) + K_o(t) \\ &= K_o(t) - x_o(t) < 0. \end{aligned}$$

Suppose  $(t_1, t_2)$  is a component of this set. Then  $t_2 = T$ , for otherwise, since  $K_o$  is decreasing on  $(t_1, t_2)$ ,  $K_o(t_2-) < 0$  and  $K_o(t_2) = K_o(t_2-) - f_o(t_2) < 0$  so that  $t_2$  would be in  $\{t: K_o(t) < 0\}$ . Thus the set must be an interval  $(b, T]$  (the same argument shows  $T$  is in

---

\* From this point on we shall think of  $f_o$  and  $K_o$  as specific functions and not as equivalence classes of functions differing on sets of measure zero.

the set) although  $K_o(T) = x_o(T) > 0$ . Thus we cannot have  $K_o \leq 0$  a.e., or  $|\{t: K_o(t) > 0\}| > 0$ .

Now as in the first minimization we can assert that

$|\{t: K_o(t) > 0\}| > a/M$ . For if this is not the case then  $f_o(t) = M$  a.e. on  $\{t: K_o(t) > 0\}$ , so that in increasing  $f_o$  to  $M$  on all this set we decrease  $K_o$  and obtain equivalent  $f_o, K_o$  for which  $f_o(t) = M$  whenever  $K_o(t) > 0$ . For this new  $K_o$ ,  $\{t: K_o(t) > 0\}$  is open; otherwise we would have  $t_n \rightarrow t$ ,  $K_o(t_n) \leq 0$ ,  $K_o(t) > 0$  and thus  $K_o(t_n) + f_o(t_n) \leq f_o(t_n) \leq M < K_o(t) + M = K_o(t) + f_o(t)$ , contradicting the continuity of  $K_o + f_o$ . Now  $K_o$  is differentiable on this set and  $K'_o(t) = f_o(t) - x_o(t) + K_o(t) = M - x_o(t) + K_o(t) > 0$ , so that  $K_o$  is strictly increasing on its components. Consequently if  $(t_1, t_2)$  is a component, then  $K_o(t_2-) > 0$  and since, by continuity,  $K_o(t_2-) + M = K_o(t) + f_o(t)$ ,  $K_o(t_2) = K_o(t_2-) + M - f_o(t_2) > 0$ , and we must have  $t_2 = T$ , and indeed  $T$  in the component. Thus  $\{t: K_o(t) > 0\} = (b, T]$ . But then  $f_o(T) = M$  and  $0 < K_o(T) = x_o(T) - M < 0$ , which is the desired contradiction.

As in the first minimization let  $\lambda$  be the supremum of all  $\mu > 0$  for which  $|\{t: K_o(t) \geq \mu\}| \geq a/M$ , so that  $|\{t: K_o(t) \geq \lambda\}| \geq a/M$  and  $|\{t: K_o(t) > \lambda\}| \leq a/M$ . As before we can modify  $f_o$ ,  $K_o$  on a set of measure zero so that  $f_o(t) = M$  whenever  $K_o(t) > \lambda$ , and by exactly the argument of the preceding paragraph we find that for the modified  $K_o$ ,  $\{t: K_o(t) > \lambda\}$  is void.

Thus we have an  $f_o$  and  $K_o$  for which  $|\{t: K_o(t) = \lambda\}| \geq a/M$  and  $K_o(t) \leq \lambda$  for all  $t$ . For this  $K_o$  and  $f_o$  we have  $f_o(t) = 0$  for all  $t$  in  $\{t: K_o(t) < \lambda\}$  outside a subset  $E$  of measure zero. Let us modify  $(f_o, K_o)$  on  $E$  to form the equivalent pair  $(f_o, K_o)$  in the following fashion: set  $E_1 = \{t: t \in E, K_o(t) + f_o(t) \leq \lambda\}$ ,

$E_2 = \{t: t \in E, x_o(t) + f_o(t) > \lambda\}$ , and

$$K_o(t) = K_o(t) + f_o(t), \quad F_o(t) = 0 \quad \text{for } t \in E_1$$

$$K_o(t) = \lambda, \quad F_o(t) = f_o(t) - (\lambda - K_o(t)) > 0 \quad \text{for } t \in E_2.$$

Then  $F_o + K_o = f_o + K_o$ ,  $K_o(t) \leq \lambda$  for all  $t$  and  $f_o(t) = 0$  whenever  $K_o(t) < \lambda$ . Omitting the bars we can now assert that  $E = \{t: K_o(t) < \lambda\}$  is open, for otherwise we have  $t_n \rightarrow t$ ,  $K_o(t_n) = \lambda$ ,  $K_o(t) < \lambda$  and  $K_o(t_n) + f_o(t_n) \geq \lambda > K_o(t) = K_o(t) + f_o(t)$ . Moreover, for any boundary point of  $E$  we have  $f_o(t) = 0$  since we have  $t_n \rightarrow t \notin E$ ,  $t_n \in E$  and thus  $K_o(t) = \lambda > K_o(t_n)$ ,

$$\begin{aligned} K_o(t_n) &= x_o(t_n) + e^{\int_{t_n}^t e^{-s} (f_o - x_o) ds} \\ &\leq K_o(t) = x_o(t) - f_o(t) + e^t \int_t^T e^{-s} (f_o - x_o) ds \end{aligned}$$

so that  $f_o(t) \leq 0$ .

Now suppose  $t_1, t_2$ ,  $t_1 < t_2$  are in the complement of  $E$  and  $(t_1, t_2) \subset E$ . Then since  $f_o(t_1) = 0$ ,  $K_o$  is continuous on  $[t_1, t_2]$ , and in  $(t_1, t_2)$ .

$$K'_o(t) = x'_o(t) + K_o(t) = K_o(t) - x_o(t)$$

$$\text{so } K''_o(t) = K'_o(t) - x'_o(t) = K_o(t) - x_o(t) - x'_o(t) = K_o(t).$$

Since  $K_o(t_1) = \lambda$ ,  $i=1,2$ , we have, on  $[t_1, t_2]$ ,

$$K_0(t) = \frac{\lambda}{1 + e^{t_2-t_1}} (e^{t-t_1} + e^{t_2-t}),$$

and, since  $x_0(t) = x_0(t_1)e^{t_1-t}$  on  $[t_1, t_2]$ ,

$$\frac{\lambda}{1 + e^{t_2-t_1}} (e^{t-t_1} + e^{t_2-t}) = x_0(t_1)e^{t_1-t} + e^t \int_t^T e^{-s} (f_0 - x_0) ds,$$

$$\frac{\lambda}{1 + e^{t_2-t_1}} (e^{-t_1} + e^{t_2-2t}) = x_0(t_1)e^{t_1-2t} + \int_t^T e^{-s} (f_0 - x_0) ds.$$

Differentiating,

$$\begin{aligned} \frac{-2\lambda}{1 + e^{t_2-t_1}} e^{t_2-2t} &= -2x_0(t_1) e^{t_1-2t} + x_0(t)e^{-t} \\ &= -2x_0(t_1) e^{t_1-2t} + x_0(t_1) e^{t_1-t} e^{-t} \\ &= -x_0(t_1) e^{t_1-2t}. \end{aligned}$$

Thus

$$x_0(t_1) = \frac{2\lambda}{1 + e^{t_2-t_1}} e^{t_2-t_1} = \frac{2\lambda}{1 + e^{t_1-t_2}} > \lambda,$$

$$x_0(t_2) = x_0(t_1)e^{t_1-t_2} = \frac{2\lambda}{1 + e^{t_2-t_1}} < \lambda.$$

These inequalities show that the components of  $E = \{t: K_0(t) < \lambda\}$  are separated by non-degenerate closed intervals contained in the complement of  $E$ . But everywhere in the complement (since  $K_0(t) \leq \lambda$  for all  $t$ )

$$\lambda = K_o(t) = x_o(t) - f_o(t) + e^t \int_t^T e^{-s} (f_o - x_o) ds$$

so that  $f_o$  is continuous and thus differentiable in such an interval and

$$\lambda e^{-t} = (x_o(t) - f_o(t)) e^{-t} + \int_t^T e^{-s} (f_o - x_o) ds,$$

$$-\lambda e^{-t} = (x'_o(t) - f'_o(t)) e^{-t} - (x_o(t) - f_o(t)) e^{-t} - e^{-t} (f_o(t) - x_o(t))$$

$$\text{and } x'_o(t) - f'_o(t) = -\lambda \text{ or } f'_o(t) = x'_o(t) + \lambda = f_o(t) - x_o(t) + \lambda.$$

$$\text{Thus } f''_o(t) = f'_o(t) - x'_o(t) = f_o(t) - x_o(t) + \lambda - x'_o(t) = \lambda > 0,$$

and since then  $f_o$  cannot be non-negative and zero at two points we must conclude that  $E$  has one component.

We may now rule out the possibility that  $E$  has a component  $(r, s)$ ,  $0 \leq r < s < T$ , for then  $\{t: K_o(t) = \lambda\}$  contains  $[s, T]$ , and since  $f''_o = \lambda$  and  $f_o(s) = 0$

$$f_o(t) = \frac{\lambda}{2} (t-s)^2 + k(t-s) \text{ for } t \geq s.$$

$$\text{Since } f'_o = x'_o + \lambda, x''_o = f''_o = \lambda \text{ and}$$

$$x_o(t) = \frac{\lambda}{2} (t-s)^2 + (k-\lambda)(t-s) + x_o(s)$$

hence

$$\lambda = K_o(T) = x_o(T) - f_o(T) = -\lambda(T-s) + x_o(s).$$

But since  $K_o(t) < K_o(s) = \lambda$  immediately to the left of  $s$  we have

$0 \leq K'_0(t) = K_0(t) - x_0(t)$  for  $t < s$  and arbitrarily close to  $s$ ,  
 so that  $\lambda = K_0(s) \geq x_0(s)$ , since  $K_0$  is continuous on  $[r, s]$ .  
 Thus since  $s < T$ ,  $\lambda > 0$ ,

$$-\lambda(T-s) + x_0(s) = \lambda > x_0(s) > x_0(s) - \lambda(T-s)$$

and we arrive at the desired contradiction.

Finally, then we know that  $K_0 = \lambda$  on an interval  $[0, b]$ ,  $f_0$  and  $x_0$  are second-degree polynomials there satisfying  $f_0(b) = 0$ ,  $x_0(0) = 1$ ,  $f'_0 = x'_0 + \lambda$ ; the polynomials

$$f_0(t) = \frac{\lambda}{2} (t-b)^2 + k(t-b)$$

$$x_0(t) = \frac{\lambda}{2} (t-b)^2 + (k-\lambda)t - \frac{\lambda}{2} b^2 + 1$$

evidently satisfy these conditions. One may now determine the unknowns  $\lambda$ ,  $b$  and  $k$  from the conditions:  $x'_0(0) = f'_0(0) = 1$ ,  $\int_0^b f_0 dt = a$ , and  $\lambda = K_0(b)$ , which yield the equations

$$-\lambda b + k - \lambda = \frac{\lambda}{2} b^2 - kb - 1,$$

$$a = \frac{1}{6} \lambda b^3 - 1/2 kb^2,$$

$$\lambda = 1/2 \left[ (k-\lambda)b - \frac{\lambda}{2} b^2 + 1 \right] (1 + e^{2B-2T}).$$

§9. The Functional  $\max |1-u|$ .

In both of the preceding problems we have found  $\min \{F(x): G(x) \leq a\}$  for two functions  $F, G$  on a set  $X$ ; such a problem has a natural dual, that of finding  $\min \{G(x): F(x) \leq b\}$ . A simple and quite useful relation between the two is furnished by the trivial

Lemma. If  $x_0$  is the unique  $x_0$  furnishing  $\min \{F(x): G(x) \leq a\} = b$  then  $x_0$  is the unique  $x$  furnishing  $\min \{G(x): F(x) \leq b\} \leq a$ .

Proof: Clearly  $G(x) \leq G(x_0) \leq a$  implies  $F(x) > F(x_0) = b$ , so that  $F(x) \leq F(x_0) = b$  implies  $G(x) > G(x_0)$ .

The usefulness of the lemma is apparent in the following problem. As before, let  $x$  be the absolutely continuous solution of  $x' = -x + f$ ,  $x(0) = 1$ , and consider minimizing

$$(1) \quad \max_t |1 - x(t)|$$

for those  $f \in L_2(0, T)$  for which  $\int_0^T f^2 dt \leq b < T$ . It is not at all apparent that there is a unique minimizing  $f$  in this case until, utilizing the lemma, we consider the dual problem. To minimize  $\int_0^T f^2 dt$  over the set  $F$  of all  $f \in L_2$  for which  $\max_t |1 - x(t)| \leq a$ , (or  $1 - a \leq x(t) \leq 1 + a$  for all  $t$ ) we are again seeking an element  $f$  of minimal norm in a strongly closed convex subset of  $L_2$ , and this element is unique. As we shall see, for  $0 \leq b \leq T$  there is a unique  $a$ ,  $0 \leq a \leq 1 - e^{-T}$ , for which the minimizing  $f$  has  $\int_0^T f^2 dt = b$  so that  $f$  minimizes (1).

Thus, we shall proceed to solve the second problem.  $f$  has the obvious property that

$$(2) \quad (f, f) \leq (f, g) \text{ for } g \in F$$

as one can see from the fact that

$$\phi(\lambda) = ||\lambda f + (1-\lambda)g||^2 = \lambda^2 (f, f) + (1-\lambda)^2 (g, g) + 2\lambda(1-\lambda) (f, g)$$

has its minimum at 1. Moreover (2) characterizes  $f$ , for if  $h \in F$  also satisfies (2) then

$$0 \leq (f-h, f-h) = [(f, f) - (f, h)] + [(h, h) - (f, h)] \leq 0 \text{ and } h = f.$$

By means of (2) we may determine  $f$  on the open set where  $1-a < x(t) < 1+a$ , or

$$(3) \quad (1-a)e^t - 1 < \int_0^t e^s f(s)ds < (1+a)e^t - 1.$$

Indeed for each component  $I$  of the set we have a constant  $c$  for which  $f(t) = ce^t$  in  $I$ . For let  $I_0$  be any closed subinterval of  $I$ ; then there is an  $\eta > 0$  for which, on  $I_0$ ,

$$(4) \quad (1-a)e^t - 1 + \eta < \int_0^t e^s f(s)ds < (1+a)e^t - 1 - \eta.$$

Now if for some  $g$  vanishing outside  $I_0$  we have  $(f, g) \neq 0$ ,  $(e^s, g) = 0$ , then  $f + \delta g$  will be in  $F$  for  $|\delta|$  small since (3) clearly holds (for  $f + \delta g$ ) for  $t$  outside  $I_0$ , and holds for  $t$  in  $I_0$  by (4). But then we may choose the sign of  $\delta$  so that  $(f, f + \delta g) = (f, f) + \delta (f, g) < (f, f)$ , a contradiction. Thus  $(e^s, g) = 0$  implies  $(f, g) = 0$  and  $f(t) = ce^t$  in  $I_0$ . Since  $I_0$  is an arbitrary closed subinterval of the open interval  $I$ , our assertion is proved.

We must now resort to another variation—that of  $T$ . Let us denote by  $F_T$  the set we have called  $F$  and by  $f_T$  the element of minimal norm in this set. We note that we may extend  $f_T$  to all of  $(0, \infty)$ , by setting  $f_T(t) = 1 - a$  for  $t > T$ , and, for the extended  $f_T$ ,  $f_T \in F_T$ , for all  $T' > 0$ , since, trivially for  $t \geq T$ ,  $(e^t x_T)' = (1-a)e^t$ ,  $e^t x_T(t) = e^T x_T(T) = (1-a)(e^t - e^T)$  and

$$x_T(t) = 1 - a + e^{T-t} [x_T(T) - (1-a)] \geq 1 - a$$

$$\leq 1 - a + x_T(T) - (1-a) = x_T(T) \leq 1 + a.$$

From the minimal property of  $f_T$ , if  $T < T'$  then

$$\int_0^T f_T^2 dt \leq \int_0^T f_{T'}^2 dt \leq \int_0^{T'} f_{T'}^2 dt \leq \int_0^{T'} f_T^2 dt = \int_0^T f_T^2 dt + (1-a)^2(T' - T).$$

It is evident from this relation that if  $T_n \rightarrow T$  then  $\{f_{T_n}\}$  is a sequence of elements of  $F_T$  whose norms tend to the minimal norm. But, as is well known,\* this implies that  $\{f_{T_n}\}$  converges strongly to  $f_T$  in  $L_2(0, T)$ .

Now if  $a \geq 1$  then  $f = 0$  is in  $F_T$ , so that we need only consider  $a < 1$ . In this case  $f = 0$  is in  $F_T$  for  $T \leq T_0 = \log 1/1-a$ , clearly, and this is not the case for  $T > T_0$ . For  $T > T_0$  we have  $x_T(1) = 1-a$ , for if this is not the case and  $t_0$  is the least  $t$  for which  $\int_t^T f_T^2 dt = 0$ , then evidently setting  $f_T = 0$  on  $(t_0 - \epsilon, t_0)$  for  $\epsilon > 0$  small yields an element of  $F_T$  of smaller norm.

\* The usual argument runs: If  $f_n \in F$ , a convex set and  $\|f_n\| \rightarrow \|f_0\| = \min_{f \in F} \|f\|$ , then  $\|f_n - f_0\|^2 + \|f_n + f_0\|^2 = 2\|f_n\|^2 + 2\|f_0\|^2$  so that

$$\|f_n - f_0\|^2 = 2\|f_n\|^2 + 2\|f_0\|^2 - 4\|\frac{f_n + f_0}{2}\|^2 \leq 2\|f_n\|^2 + 2\|f_0\|^2 - 4\|f_0\|^2 \rightarrow 0.$$

As a consequence of these facts we can assert that  $x_T$  is non-increasing for all  $T > 0$ . Obviously this is the case for  $T \leq T_0$ , and if this is not true for some  $x_T$  then we have  $t, t'$ ,  $0 \leq t < t' \leq T$  for which  $x_T(t) < x_T(t')$ ; consequently this must hold for two points  $t, t'$  in the open set where  $1 - a < x_T(t) < 1 + a$ , indeed for two points in the same component I of this set. But on I,  $f(t) = ce^t$ ,  $x'_T = -x_T + ce^t$  so that  $c > 0$ ; consequently  $x_T$  can have no maximum on I since at a maximum we would have  $0 = x'_T = -x_T + ce^t$

$$0 \geq x''_T = -x'_T + ce^t = ce^t > 0.$$

Inasmuch as  $x_T(T) = 1 - a$ ,  $x_T$  must then assume the value  $1 + a$  at the endpoint of I. Thus for those  $T$  for which  $x_T$  is not non-increasing  $\max x_T = 1 + a$ , while on the complementary set  $\max x_T = 1$ .

Now as  $T_n \rightarrow T$  we must clearly have  $x_{T_n} \rightarrow x_T$  uniformly on any finite interval  $(0, K)$  for

$$\begin{aligned} |x_{T_n}(t) - x_T(t)| &= e^{-t} \left| \int_0^t e^s (f_{T_n} - f_T) ds \right| \\ &\leq e^{-t} \left( \frac{e^{2t} - 1}{2} \right)^{1/2} \left( \int_0^t (f_{T_n} - f_T)^2 ds \right)^{1/2} \\ &\leq \left( \frac{e^{2K} - 1}{2} \right)^{1/2} \left( \int_0^{T^*_n} (f_{T_n} - f_T)^2 ds \right)^{1/2} \end{aligned}$$

where  $T^*_n$  is the larger of  $T$  and  $T_n$ . Therefore, the set of  $T > 0$  where  $x_T$  is non-increasing and its complement are closed subsets of  $(0, \infty)$ . Since the former is non-void the latter is void, and  $x_T$  is non-increasing for all  $T > 0$ .

The form of  $f_T$  for  $T > T_0$  is now clear:  $f_T(t) = ce^t$ ,  $0 \leq t \leq \xi$ ;  $f_T(t) = 1 - \alpha$ ;  $\xi < t \leq T$ . Let us set  $\alpha = 1 - a$ . The relationship between  $\xi$  and  $c$  is found from the solution of  $x' = -x + ce^t$ ,  $x(0) = 1$ , that is

$$(5) \quad x(t) = (1 - \frac{c}{2})e^{-t} + \frac{c}{2}e^t,$$

by virtue of the fact that  $x(\xi) = \alpha$  or

$$(6) \quad (1 - \frac{c}{2})e^{-\xi} + \frac{c}{2}e^\xi = \alpha.$$

If we denote by  $\phi(c)$  the value of  $\int_0^T f^2 dt$  where

$$f(t) = \begin{cases} ce^t, & 0 \leq t \leq \xi(c) \\ \alpha, & \xi(c) < t \leq T \end{cases},$$

then

$$(7) \quad \phi(c) = c^2 \frac{e^{2\xi} - 1}{2} + \alpha^2 (T - \xi).$$

To find  $f_T$  we must now minimize  $\phi(c)$  over all  $c$ , where  $\xi$  and  $c$  are connected by (6), and  $c$  must satisfy the additional constraints that  $x$  be decreasing and  $0 \leq \xi(c) \leq T$ .

Clearly taking  $c < 0$  is inferior to taking  $c = 0$ , and  $c > 1$  yields an increasing  $x$  so that we need only consider  $0 \leq c \leq 1$ . Differentiating (6) we obtain

$$\begin{aligned} 0 &= \frac{1}{2} (e^\xi - e^{-\xi}) + \left( \frac{c}{2}e^\xi - (1 - \frac{c}{2})e^{-\xi} \right) \frac{d\xi}{dc} \\ &= \frac{1}{2} (e^\xi - e^{-\xi}) + (ce^\xi - \alpha) \frac{d\xi}{dc} \end{aligned}$$

while differentiation of (7) yields

$$\begin{aligned}
 \phi'(c) &= c(e^{2\xi} - 1) + (c^2 e^{2\xi} - \alpha^2) \frac{d\xi}{dc} \\
 &= c(e^{2\xi} - 1) - \frac{1}{2} (e^\xi - e^{-\xi})(ce^\xi + \alpha) \\
 &= (e^\xi - e^{-\xi})(ce^\xi - \frac{1}{2} ce^\xi - \frac{1}{2} \alpha) \\
 &= \frac{1}{2} (e^\xi - e^{-\xi})(ce^\xi - \alpha).
 \end{aligned}$$

Moreover, if we note that our non-increasing solution  $x_T$  is, from (5), a convex combination of two functions, it is evident from (6) that  $\xi$  increases with  $c$ . Since  $\phi'(c) < 0$  for  $ce^\xi < \alpha$ ,  $\phi'(c) > 0$  for  $ce^\xi > \alpha$ , it follows that the value  $c^*$  provided by  $c^* e^{\xi^*} = \alpha$  will provide our minimum if  $\xi^* \leq T$ ; otherwise (since  $\xi$  increases with  $c$ )  $\phi'(c) < 0$  for all  $\xi < T$  and we must take  $\xi = T$ . Now from (6),  $(1 - \frac{c^*}{2})e^{-\xi^*} = \frac{1}{2}\alpha$  so

$$(1 - \frac{c^*}{2}) \frac{2c^*}{\alpha} = \alpha,$$

$$c^{*2} - 2c^* + \alpha^2 = 0$$

and  $c^* = 1 - \sqrt{1 - \alpha^2}$ . Thus  $\xi^* = \log \frac{\alpha}{1 - \sqrt{1 - \alpha^2}}$  and we have

$$r_T(t) = \frac{2(\alpha - e^{-T})}{e^T - e^{-T}} e^t, \quad 0 \leq t \leq T$$

for  $T_0 = \log \frac{1}{\alpha} \leq T \leq \log \frac{\alpha}{1 - \sqrt{1 - \alpha^2}}$ , and

$$r_T(t) = \begin{cases} (1 - \sqrt{1 - \alpha^2})e^t, & 0 \leq t \leq \log \frac{\alpha}{1 - \sqrt{1 - \alpha^2}} \\ \alpha & t \geq \log \frac{\alpha}{1 - \sqrt{1 - \alpha^2}} \end{cases}$$

for all larger  $T$ .

Consider now the original problem, that of finding an  $f$  with  $\int_0^T f^2 dt \leq b$  for which  $\max_{0 \leq t \leq T} |1 - x(t)|$  is a maximum. From

the solution  $f$  to the dual problem just obtained we see that  $\int_0^T f^2 dt$  is a continuous function of  $a$ , and for  $0 \leq a \leq 1 - e^{-T}$  it is easily seen to be strictly decreasing with values  $T$  and 0 for  $a = 0$  and  $a = 1 - e^{-T}$ . Thus for each  $b$  in the range  $0 \leq b \leq T$  we have an  $a$ ,  $0 \leq a \leq 1 - e^{-T}$  for which the corresponding  $f$  provides  $\min \max_t |1 - x(t)|$ .

#### REFERENCES

1. Bellman, R. "On the Theory of Dynamic Programming," Proc. Nat. Acad. Sci., Vol. 38 (1952), pp. 716-719.
2. Bellman, R., Glicksberg, I., Gross, O. "On Some Variational Problems Occurring in the Theory of Dynamic Programming," Proc. Nat. Acad. Sci., Vol. 39 (1953).

bjc